Policy for Dataset Preparation

Note that the following policy is an update from the 10/25/2002 policy.

For contract-supported clinical trials and epidemiology studies: Requirements for preparation of data sets have been modified to shorten the timeline and expand the data to be included, as described below. These changes will be effective with contracts awarded on or after October 1, 2005.

For grant-supported new and competing applications for selected epidemiology studies and clinical trials: Applications received on or after October 1, 2005 will be expected to include provisions for submission of data sets as described below. Applicants are expected to include the costs of data set preparation, with appropriate justification, in their budget requests. Funds awarded for data set preparation will be restricted for use solely for that purpose and only upon release by the NHLBI.

In general the following types of studies will be included under this policy:

- Clinical trials and epidemiology studies that are supported by the U01 (cooperative agreement) mechanism AND have 500 or more participants
- Trials or studies requesting $500,000 direct costs or more in any one year and identified as being of high programmatic interest to the NHLBI, as indicated in the Institute's letter of agreement to accept assignment of the application
- Ancillary studies based on clinical trials or epidemiology studies that are required by this policy to provide NHLBI data sets.

Requests for exceptions to these guidelines will be considered by the NHLBI if adequately justified. Examples of adequate justification include: unavoidable and unanticipated delays in making data available within the parent study for analysis; presence of provisions in informed consent prohibiting data set release; evidence of unacceptability of data set release to communities under study; measurements on too small a subset of participants to be of scientific value. All such requests should be addressed to the Director of the Program Division funding the award.

Policies for data sharing from studies of American Indian or Alaska Native tribes and other sovereign entities will be developed with them and will be provided as available at a later date.

I. Introduction

The National Heart, Lung, and Blood Institute (NHLBI) has supported data collection from participants in numerous clinical trials and epidemiologic studies. These data from well-characterized population samples constitute an important scientific resource. It is the view of the NHLBI that their full value can only be realized if they are made available, under appropriate terms and conditions consistent with the informed consent provided by individual participants, in a timely manner to the largest possible number of qualified investigators.

Under no circumstances will data relating to an individual be distributed in any way that is inconsistent with his or her informed consent. Data sets without an informed consent permitting use by non-study researchers will only be released if the requester's IRB has approved a waiver of informed consent based on minimal risk to the participants [see Institutional Review Board section].

Data sets distributed under this policy include only data with personal identifiers and other variables that might enable individual participants to be identified, such as outliers, dates, and study sites, removed or otherwise modified. Because it may still be possible to combine the data with other publicly available data and thereby determine with reasonable certainty the identity of individual participants, these data sets are not truly anonymous. They are, therefore, only provided to investigators who agree in advance to adhere to established policies for distribution.

Data sets are available for NHLBI studies supported by contract and for selected studies supported by cooperative agreements or other grants. However, data will not be provided if the Institute deems them to be unreliable or invalid. All proposed data exclusions must be strongly justified and whether proposed by the study investigators or Institute staff, each one must be reviewed and approved by the director of the NHLBI program division that sponsored the study.

II. Definitions

Data - Information collected and recorded from study participants through periodic examinations and follow-up contacts, not to include original specimens or images.

Commercial purpose - Data will be considered as being for a commercial purpose if they are to be used by an investigator who is an employee of a for-profit organization, if they are to be used by an investigator to satisfy a contractual relationship with a for-profit organization, or if they are to be used by an investigator as the basis for a consulting relationship with a for-profit organization. Data will also be considered as being for a commercial purpose if the investigator(s) take any affirmative steps to facilitate commercial use of results derived from the data.

Non-Commercial Purpose Data Set - A data set consisting of all records except those for participants who requested that their data not be shared beyond the initial study investigators.

Commercial Purpose Data Set - A data set consisting of all records except those for participants who requested that their data not be shared beyond the initial study investigators or used for commercial purposes.

Non-Commercial Purpose Pedigree/Genetic Data Set - A pedigree/genetic data set consisting of all pedigree and genetic data except those for participants who requested that their data not be shared beyond the initial study investigators.

Commercial Purpose Pedigree/Genetic Data Set - A pedigree/genetic data set consisting of all pedigree and genetic data except those for participants who requested that their data not be shared beyond the initial study investigators or used for commercial purposes.

III. Data Set Requests

1. **Responsibilities of Study Investigators in Preparing Data Sets**

   Investigators in NHLBI studies covered by this policy are required as part of the terms and conditions of their awards to prepare and deliver to the NHLBI data sets that satisfy NHLBI requirements. Included among them are documentation, elimination of personal identifiers, and modification of other data elements so as to reduce the likelihood that any individual participant can be identified.

   Two data sets, i.e., a Non-Commercial Purpose Data Set and a Commercial Purpose Data Set, and, if applicable, two pedigree/genetic data sets, i.e., a Non-Commercial Purpose Pedigree/Genetic Data Set and a Commercial Purpose Pedigree/Genetic Data Set, and associated documentation, must be provided in electronic form to the Institute. In addition, investigators must provide the Institute with two separate lists of participant identification numbers, one consisting of those participants who asked that their data not to be shared beyond the initial study investigators and the other of those participants who asked that their data not be used for commercial purposes.

   Investigators in ancillary studies based on ongoing (parent) studies that are required by this policy to produce data sets must submit ancillary study data to the NHLBI through the parent study Coordinating Center or data submission process established by the parent study. Ancillary studies conducted on small subsets of a study sample may be appropriate for exclusion from data sets; requests for their exclusion should be justified and addressed as described in the Introduction above.

   1. *Documentation* - Documentation for data sets must be comprehensive and sufficiently clear to enable investigators who are not familiar with a data set to use it. The documentation must include data collection forms, study procedures and protocols, descriptions of all variable recoding performed, and a list of major study publications.

      In addition, a summary documentation file, usually called a "readme" file, is required. It must provide a complete overview of the data and a description of their use for investigators who are not familiar with the data set. It must also contain a brief description of the study (including a general orientation to the study, its components, and its examination and follow-up timeline), a listing of all files being provided, a description of system requirements, a generation program code for installing a SAS file from the SAS export data file, and a frequency distribution for selected key variables.

      Selected Documentation will be used to describe the study on the Data Repository website. Examples include Forms, Data Dictionaries, Descriptive Statistics, and the Study Protocol. These documents will need to be accessible to those with disabilities according to section 508 of the Rehabilitation Act. The HHS maintains

a website devoted to 508 issues with links to resources on creating and checking accessibility.

2. *Data Storage and Format* - The data are to be stored on a CD ROM unless the investigators and the NHLBI mutually agree upon an alternative storage medium. Both the comprehensive documentation and the summary documentation must be prepared in a consistent format, either as a Word Perfect, MS Word, ASCII, or portable document format (PDF) file and included on the same storage medium as the data set. To ensure access by users with disabilities, all PDF files must be created in Adobe Acrobat version 5.0 or higher. Documentation that is not available in electronic form, such as data collection forms, should be scanned into a graphics file, converted to a PDF file using Adobe Acrobat version 5.0 or higher, and saved on the same medium as the data set. Pedigree data should be provided in a format readable by standard genetic analysis programs such as SAGE and SOLAR, with one individual's data per line beginning with pedigree identifier, individual's id, father's id, mother's id, and individual's sex.
3. *Content of NHLBI Data* - In addition to summary information, data sets also include for each participant those raw data elements (e.g., food item data or individual electrocardiographic lead scores) *that have not otherwise been processed into summary information*.
    1. *Clinical Trials* - included are baseline, interim visit, ancillary data, and outcome data, along with laboratory measurements not otherwise summarized.
    2. *Observational Epidemiology Studies* - included are all of the examination data obtained in each examination cycle, ancillary data, and/or all of the follow-up information available up to the last follow-up cycle cutoff date
4. *Timing of Release of NHLBI Data*
    1. *Clinical Trials* - Data are prepared by the study coordinating center and sent to the NHLBI after publication of the primary clinical trial results. They are available for release once they are received and checked by the NHLBI. The data sets must be submitted to the NHLBI no later than 3 years after the final visit of the participants to their clinical trial sites or 2 years after the main paper of the trial has been published, whichever comes first.
    2. *Observational Epidemiology Studies* - Epidemiology studies typically have an examination component and a mortality/morbidity follow-up component. Data from each cycle of an examination or follow-up component are prepared by the study coordinating center and sent to the NHLBI for distribution as a data set no later than 3 years after the completion of each examination or follow-up cycle or 2 years after the baseline, follow-up, genetic, ancillary study, or other data set is finalized within the study for analysis for use in publication, whichever comes first.
    3. *Ancillary Studies* - In those cases in which the timeline for an ancillary study differs from that of its parent study, the release date will relate to the timeline of the ancillary study.

IV. Procedures for Protection of Privacy for NHLBI Data Sets

1. **Institute Review and Approval of Data Set Preparation**

   The NHLBI requires that the data be provided in a manner that protects the privacy of study participants. The Institute requires appropriate documentation of the steps taken to protect their privacy in preparing a data set. A summary of all proposed modifications and deletions to be made to a data set in preparing it must be submitted to and approved by the director of the division that sponsored the study prior to their implementation.

2. **Guidelines for NHLBI Data Set Preparation**

   The following guidelines provide a framework for decision-making regarding preparation of data sets:

   1. All data for participants who refused to permit sharing their data with other researchers must be deleted from the Non-Commercial Purpose Data Set.
   2. All data for participants who only refused to permit sharing their data for commercial purposes must also be deleted from the Commercial Purpose Data Set.
   3. Participant identifiers:
      1. Obvious identifiers (e.g., name, addresses, social security numbers, place of birth, city of birth, contact data) must be deleted.
      2. New identification numbers must replace original identification numbers. Codes linking the new and original data should be sent to the NHLBI in a separate file, not included on the CD ROM, so that linkage may be made if necessary for future research.
      3. Variables that might lead to the identification of participants and of centers in multicenter studies, or variables that are sensitive, inaccurate, or of limited scientific utility:
         1. Clinical center identifier -- In trials or studies that have only a few centers and relatively few participants per center, the data set should not contain center identifiers. In trials that have either many centers or a large number of participants per center, the data may offer little possibility of identifying individuals. For them, the investigators and the NHLBI will determine whether to include them on a case-by-case basis.
         2. Interviewer or technician identification numbers must be recoded or deleted.
         3. Sensitive data, including illicit drug use, risky behaviors (e.g., carrying a gun or exhibiting violent behavior), sexual behaviors, and selected medical conditions (e.g., alcoholism, HIV/AIDS) must be deleted.
         4. Regional variables with little or no variation within a center because they could be used to identify that center must be deleted

5. Unedited, verbatim responses that are stored as text data (e.g., specified in "other" category) must be deleted
6. Pedigree and genetic data will be distributed in separate data sets only to investigators specifically requesting them. Genotyping data for any person in whom potential pedigree errors are detected must be deleted.
7. Dates: All dates should be coded relative to a specific reference point (e.g., date of randomization or study entry). This provides privacy protection for individuals known to be in a study who are known to have had some significant event (e.g., a myocardial infarction) on a particular date.
8. Variables with low frequencies for some values, that might be used to identify participants, may be recoded. These might include:
    1. Socioeconomic and demographic data (e.g., marital status, occupation, income, education, language, number of years married).
    2. Household and family composition (e.g., number in household, number of siblings or children, ages of children or step-children, number of brothers and sisters, relationships, spouse in study).
    3. Numbers of pregnancies, births, or multiple children within a birth.
    4. Anthropometry measures (e.g., height, weight, waist girth, hip girth, body mass index).
    5. Physical characteristics (e.g., missing limbs).
    6. Detailed medication, hospitalization, and cause of death codes, especially those related to sensitive medical conditions as listed above, such as HIV/AIDS or psychiatric disorders.
    7. Prior medical conditions with low frequency (e.g., group specific cancers into broader categories) and related questions such as age at diagnosis and current status
    8. Parent and sibling medical history (e.g., parents' ages at death).
    9. Race/ethnicity and sex information when very few participants are in certain groups or cells.
    10. Polychotomous variables: values or groups should be collapsed so as to ensure a minimum number of participants (e.g., at least 20) for each value within each race-sex cell.
    11. Continuous variables: distributions should be truncated if needed to ensure that a minimum number of participants (e.g., at least 20) have the same highest and lowest values in each race-sex cell.
    12. Dichotomous variables: data should either be grouped with other related variables so as to ensure a minimum number

of participants (e.g., at least 20) in each race-sex cell or deleted

13. The investigators may realize that other variables may make it easy to identify individuals. All such variables should be recoded or removed. The NHLBI program officer or project administrator should be consulted concerning such variables.